WEEK 3

Exploratory Data Analysis

Module 3 Introduction

Exploratory Data Analysis (EDA)

- Preliminary step in data analysis to:
  - Summarize main characteristics of the data
  - Gain better understanding of the data set
  - Uncover relationships between variables
  - Extract important variables

- Question:

  "What are the characteristics which have the most impact on the car price?"

Learning Objectives

In this lesson you will learn about:

- Descriptive Statistics
- GroupBy
- Correlation
- Correlation - Statistics

Descriptive Statistics

Descriptive Statistics

- Describe basic features of data
- Giving short summaries about the sample and measures of the data

Descriptive Statistics- Describe()

- Summarize statistics using pandas **describe()** method

```
df.describe()
```

	Unnamed: 0	symboling	normalized-losses	wheel-base	length	width	height	curb-weight	engine-size	bore	stroke
count	201.000000	201.000000	164.000000	201.000000	201.000000	201.000000	201.000000	201.000000	201.000000	201.000000	201.000000
mean	100.000000	0.840796	122.000000	98.797015	174.200995	65.889055	53.766667	2555.666667	126.875622	3.319154	3.256766
std	58.167861	1.254802	35.442168	6.066366	12.322175	2.101471	2.447822	517.296727	41.546834	0.280130	0.316049
min	0.000000	-2.000000	65.000000	86.600000	141.100000	60.300000	47.800000	1488.000000	61.000000	2.540000	2.070000
25%	50.000000	0.000000	NaN	94.500000	166.800000	64.100000	52.000000	2169.000000	98.000000	3.150000	3.110000
50%	100.000000	1.000000	NaN	97.000000	173.200000	65.500000	54.100000	2414.000000	120.000000	3.310000	3.290000
75%	150.000000	2.000000	NaN	102.400000	183.500000	66.600000	55.500000	2926.000000	141.000000	3.580000	3.410000
max	200.000000	3.000000	256.000000	120.900000	208.100000	72.000000	59.800000	4066.000000	326.000000	3.940000	4.170000

Question

What happens if the method describe is applied to a dataframe with NaN values

- ○ an error will occur
- ○ all the statistics calculated using NaN values will also be NaN
- ◉ NaN values will be excluded

Correct
correct

Skip Continue

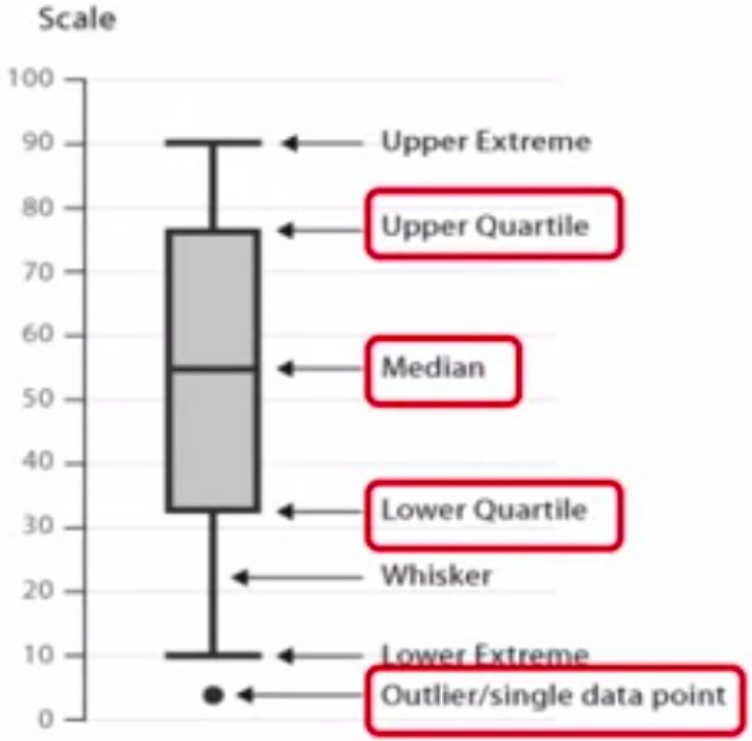
Descriptive Statistics - Value_Counts()

- summarize the categorical data is by using the **value_counts()** method

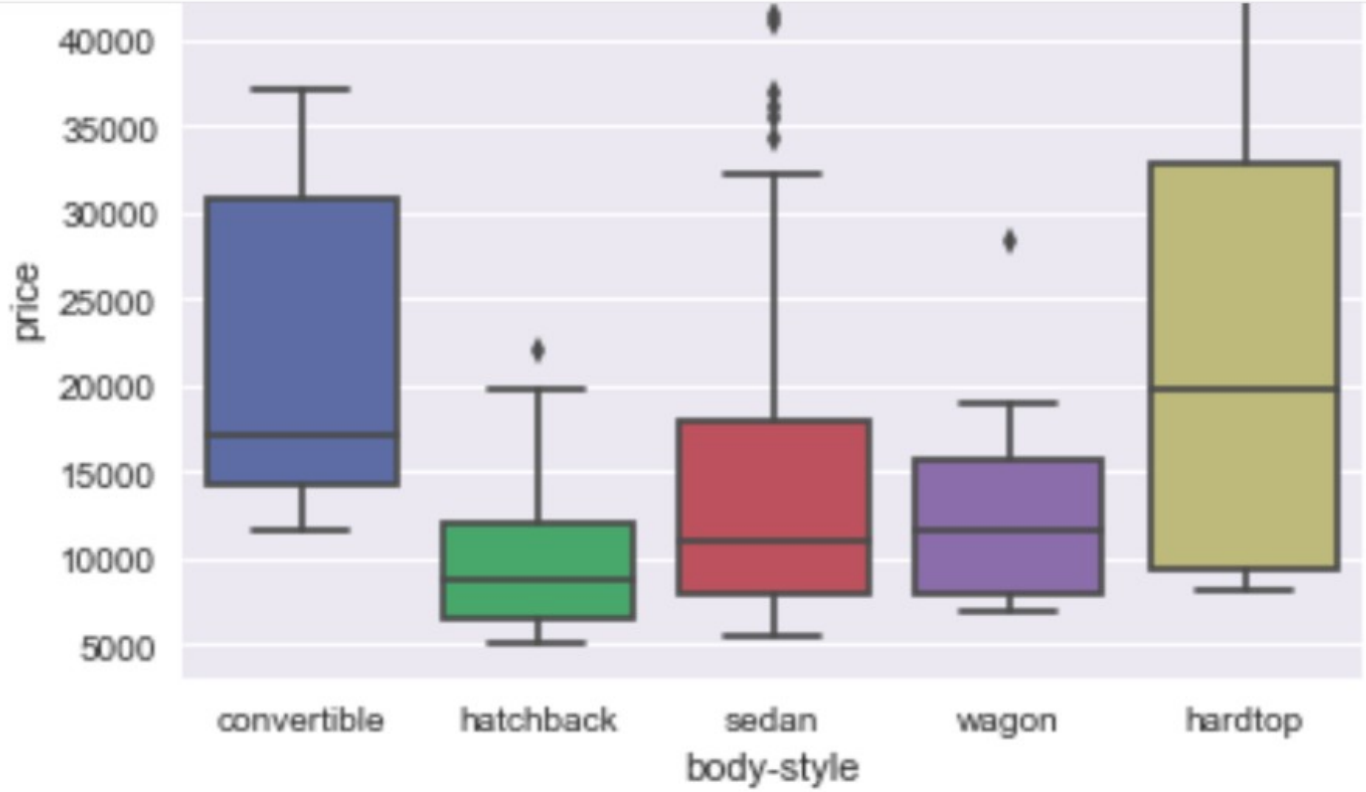
```
drive_wheels_counts=df["drive-wheels"].value_counts().to_frame()

drive_wheels_counts.rename(columns={'drive-wheels':'value_counts'}, inplace=True)
drive_wheels_counts
```

	value_counts
drive-wheels	
fwd	118
rwd	75
4wd	8

Descriptive Statistics - Box Plots



Question



Skip Continue

Question

```
sns.boxplot(x="body-style", y="price", data=df)
```

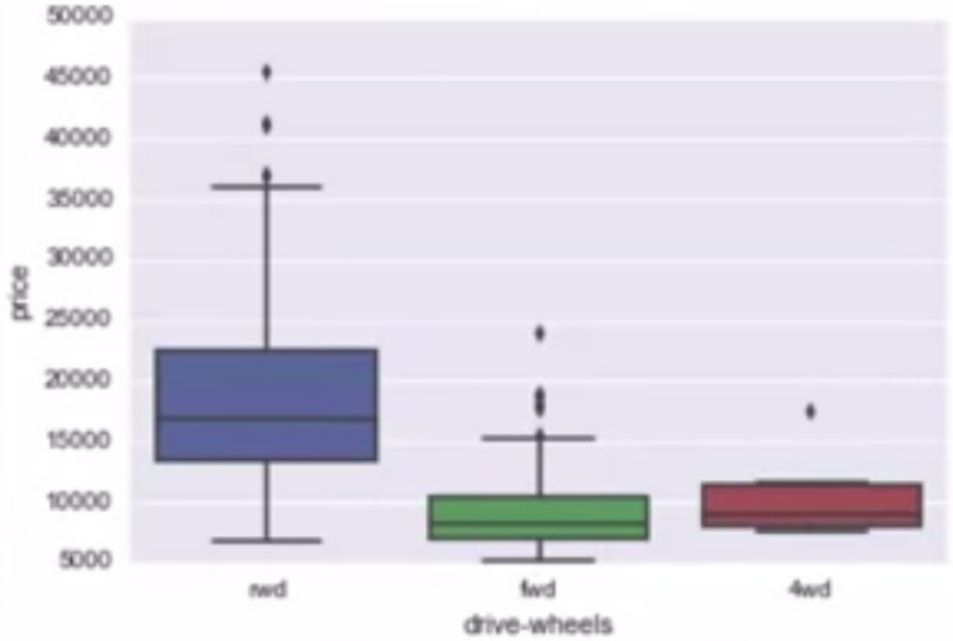
```
sns.boxplot(x="engine-location", y="price", data=df)
```

```
sns.boxplot(x="drive-wheels", y="price", data=df)
```

Skip Continue

Box Plot - Example

```
sns.boxplot(x= "drive-wheels", y= "price",data=df)
```



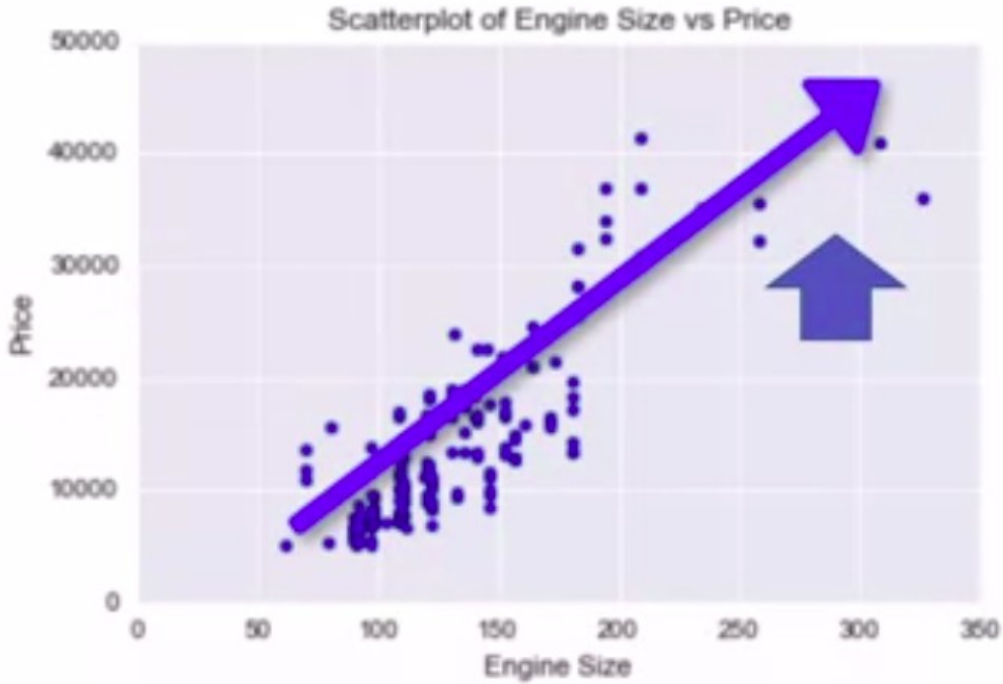
Descriptive Statistics - Scatter Plot

- Each observation represented as a point.
- Scatter plot Show the relationship between two variables.

1. Predictor/independent variables on x-axis.
2. Target/dependent variables on y-axis.

Scatterplot - Example

```
y=df["price"]
x=df["engine-size"]
plt.scatter(x,y)

plt.title("Scatterplot of Engine Size vs Price")
plt.xlabel("Engine Size")
plt.ylabel("Price")
```



Back

Practice Quiz: Descriptive Statistics

Practice Quiz • 3 min • 1 total point

Congratulations! You passed!

Grade received 100% **To pass** 50% or higher

Go to next item

1. What plot would you see after running the following lines of code?

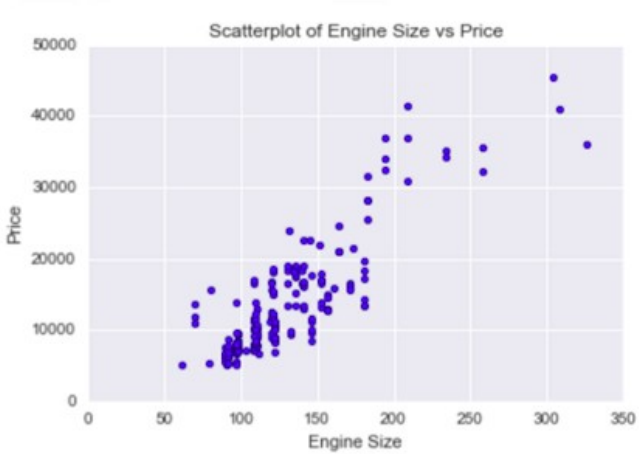
1 / 1 point

```
x=df["engine-size"]
y=df["price"]
plt.scatter(x,y)
plt.title("Scatterplot of Engine Size vs Price")
plt.xlabel("Engine Size")
plt.ylabel("Price")

```

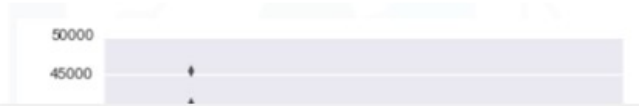
Back

Practice Quiz: Descriptive Statistics

Practice Quiz • 3 min • 1 total point



a

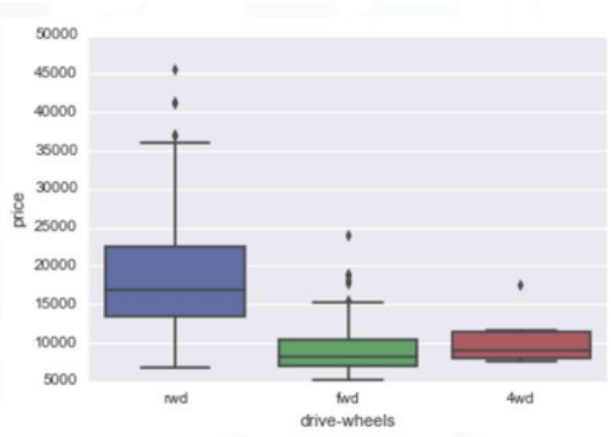


Back

Practice Quiz: Descriptive Statistics

Practice Quiz • 3 min • 1 total point

a



b

Correct

correct

GroupBy in Python

Grouping data

- Use Panda **dataframe. Groupby()** method:
  - Can be applied on categorical variables
  - Group data into categories
  - Single or multiple variables

Groupby()- Example

```
df_test = df[['drive-wheels', 'body-style', 'price']]
df_grp = df_test.groupby(['drive-wheels', 'body-style'], as_index=False).mean()
df_grp
```

	drive-wheels	body-style	price
0	4wd	convertible	20239.229524
1	4wd	sedan	12647.333333
2	4wd	wagon	9095.750000
3	fwd	convertible	11595.000000
4	fwd	hardtop	8249.000000
5	fwd	hatchback	8396.387755
6	fwd	sedan	9811.800000
7	fwd	wagon	9997.333333
8	rwd	convertible	23949.600000
9	rwd	hardtop	24202.714286
10	rwd	hatchback	14337.777778
11	rwd	sedan	21711.833333
12	rwd	wagon	16994.222222

Question

How would you use the **groupby** function to find the average "price" of each car based on "body-style" ?

- [x]
```
df[['price','body-style']].groupby(['body-style'],as_index= False).mean()
```

- [ ]
```
df.groupby(['price" ],as_index= False).mean()
```

- [ ]
```
mean(df.groupby(['price','body-style'],as_index= False))
```

Skip Continue

Pandas method - Pivot()

- One variable displayed along the columns and the other variable displayed along the rows.

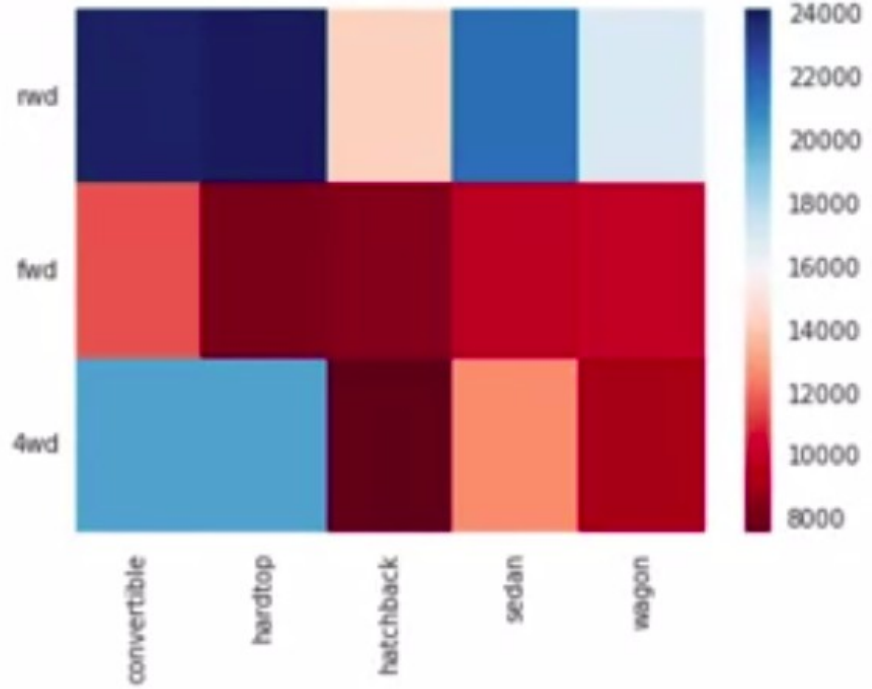
```
df_pivot = df_grp.pivot(index= 'drive-wheels', columns='body-style')
```

	price				
body-style	**convertible**	**hardtop**	**hatchback**	**sedan**	**wagon**
drive-wheels					
4wd	20239.229524	20239.229524	7603.000000	12647.333333	9095.750000
fwd	11595.000000	8249.000000	8396.387755	9811.800000	9997.333333
rwd	23949.600000	24202.714286	14337.777778	21711.833333	16994.222222

Heatmap

- Plot target variable over multiple variables

```
plt.pcolor(df_pivot, cmap='RdBu')
plt.colorbar()
plt.show()
```



Back

Practice Quiz: GroupBy in Python

Practice Quiz • 3 min • 1 total point

1. Which of the following tables representing number of drive wheels, body style and price is a Pivot Table?

1 / 1 point

	price				
body-style	**convertible**	**hardtop**	**hatchback**	**sedan**	**wagon**
drive-wheels					
4wd	20239.229524	20239.229524	7603.000000	12647.333333	9095.750000
fwd	11595.000000	8249.000000	8396.387755	9811.800000	9997.333333
rwd	23949.600000	24202.714286	14337.777778	21711.833333	16994.222222

a)

Back

Practice Quiz: GroupBy in Python

Practice Quiz • 3 min • 1 total point

	drive-wheels	body-style	price
0	4wd	hatchback	7603.000000
1	4wd	sedan	12647.333333
2	4wd	wagon	9095.750000
3	fwd	convertible	11595.000000
4	fwd	hardtop	8249.000000
5	fwd	hatchback	8396.387755
6	fwd	sedan	9811.800000

Correlation

Correlation

What is Correlation?

- Measures to what extent different variables are interdependent.

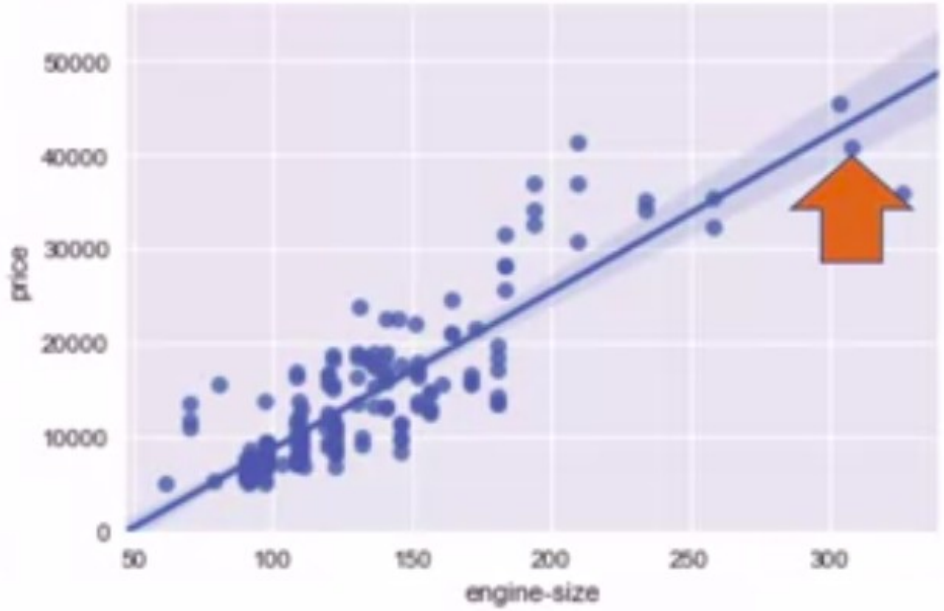
- For example:
  - Lung cancer → Smoking
  - Rain → Umbrella

- Correlation doesn't imply causation.

Correlation - Positive Linear Relationship

- Correlation between two features (engine-size and price).

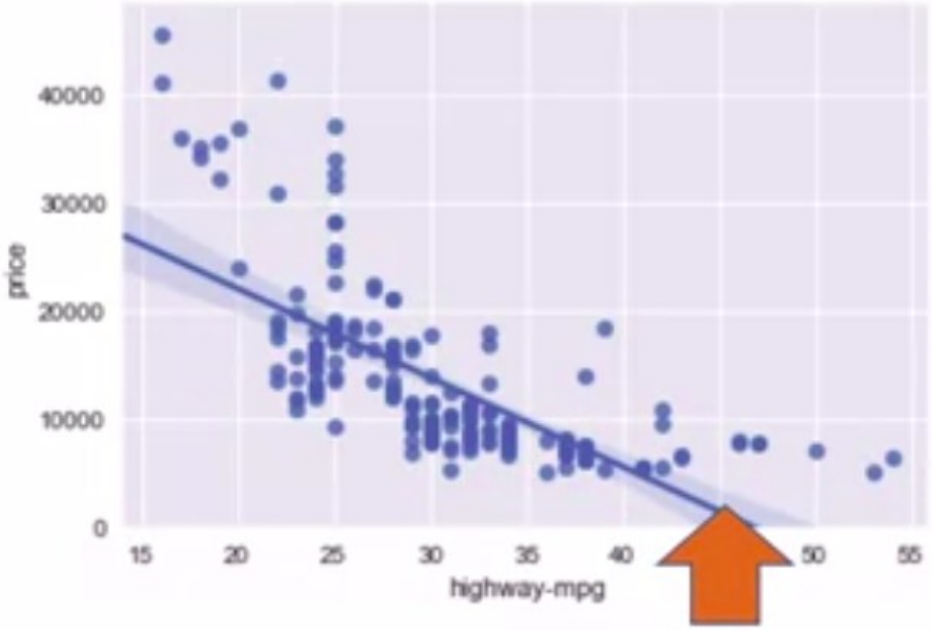
```
sns.regplot(x="engine-size", y="price", data=df)
plt.ylim(0,)
```



Correlation - Negative Linear Relationship

- Correlation between two features (highway-mpg and price).

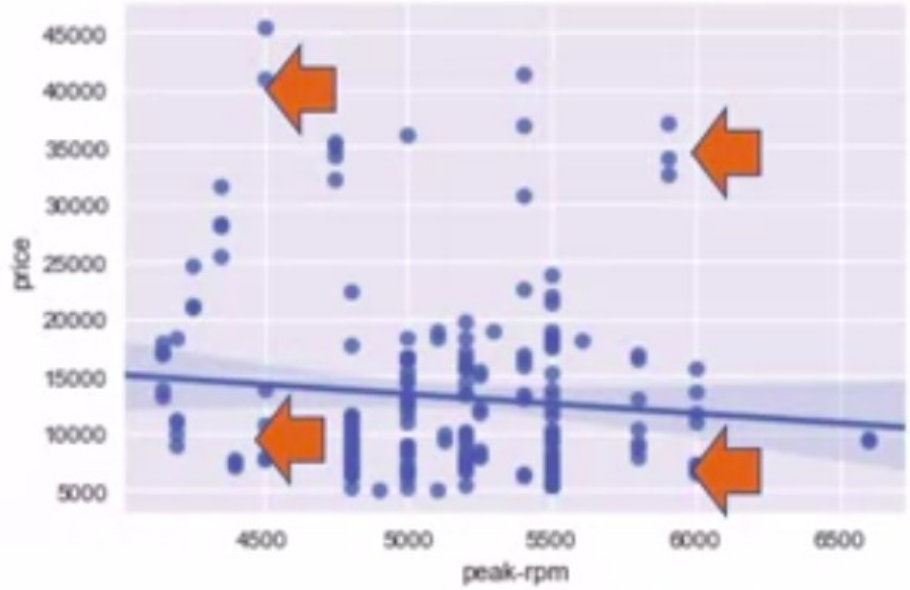
```
sns.regplot(x="highway-mpg", y="price", data=df)
plt.ylim(0,)
```



Correlation - Negative Linear Relationship

- Weak correlation between two features (peak-rpm and price).

```
sns.regplot(x="peak-rpm", y="price", data=df)
plt.ylim(0,)
```



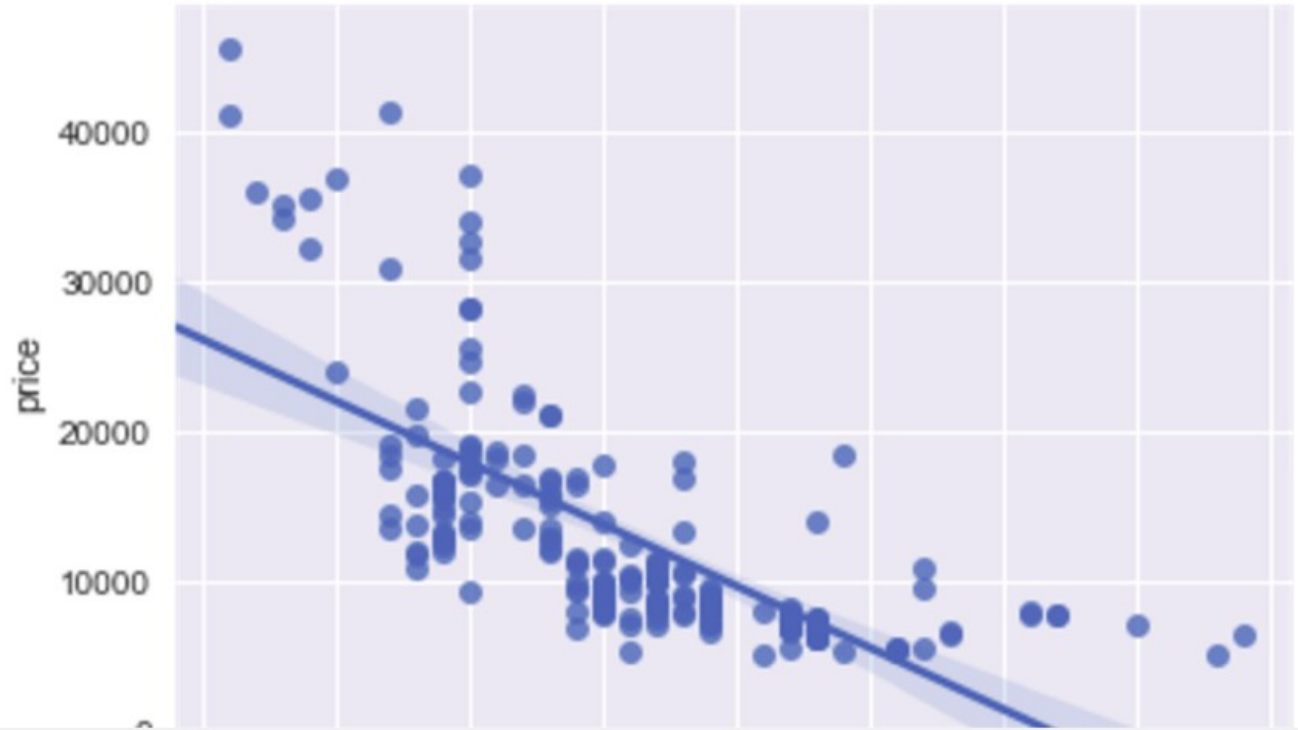
Back

Correlation

Practice Quiz • 3 min • 1 total point

1. Select the scatter plot with weak correlation:

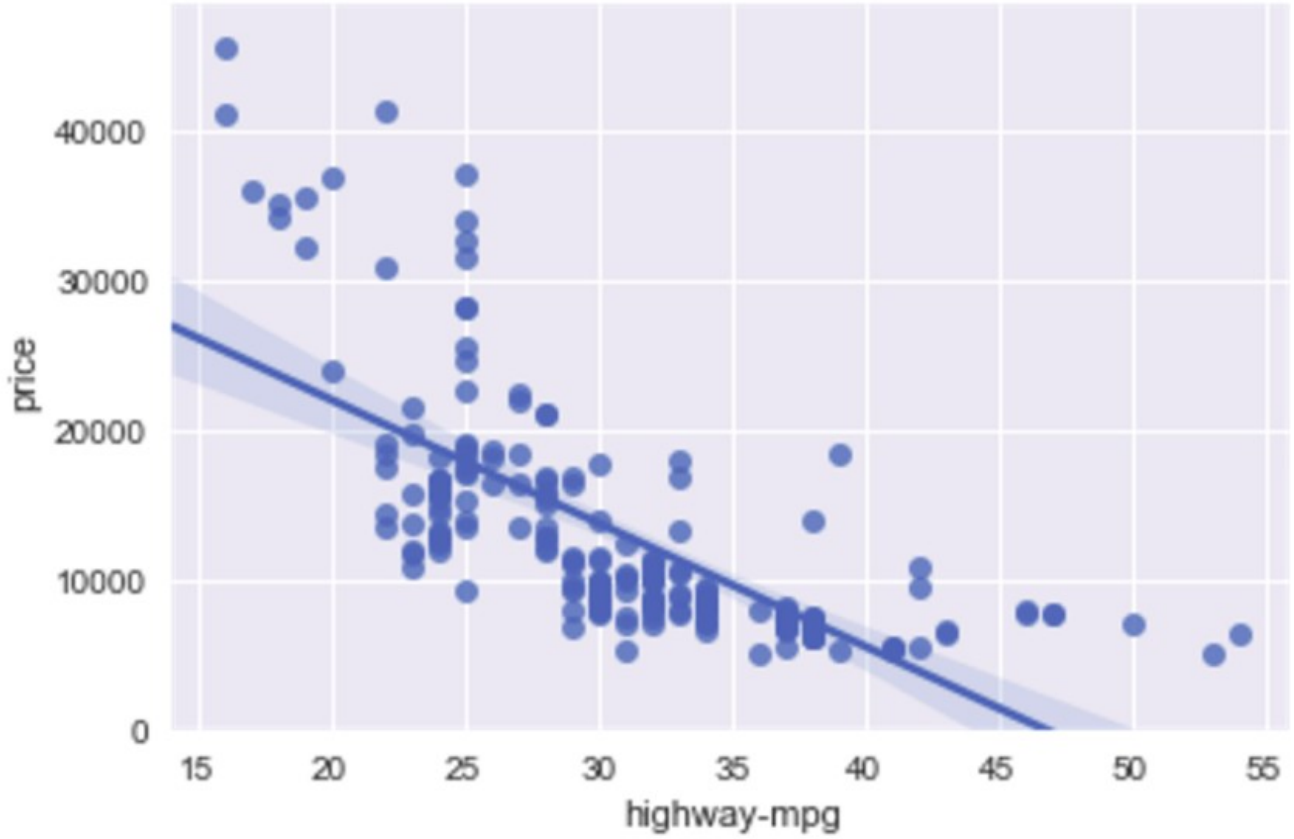
1 / 1 point



Back

Correlation

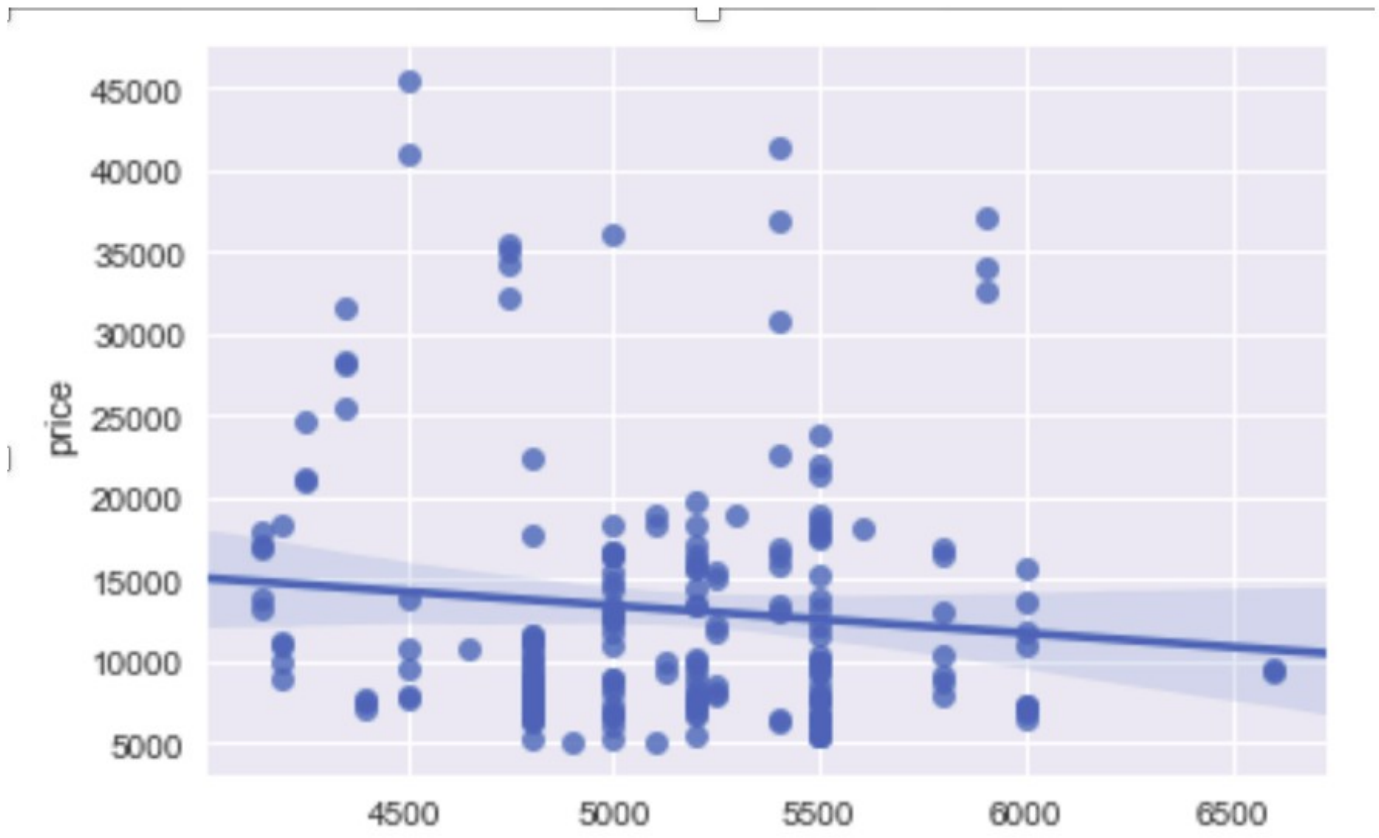
Practice Quiz • 3 min • 1 total point



Back

Correlation

Practice Quiz • 3 min • 1 total point



Correlation - Statistics

Pearson Correlation

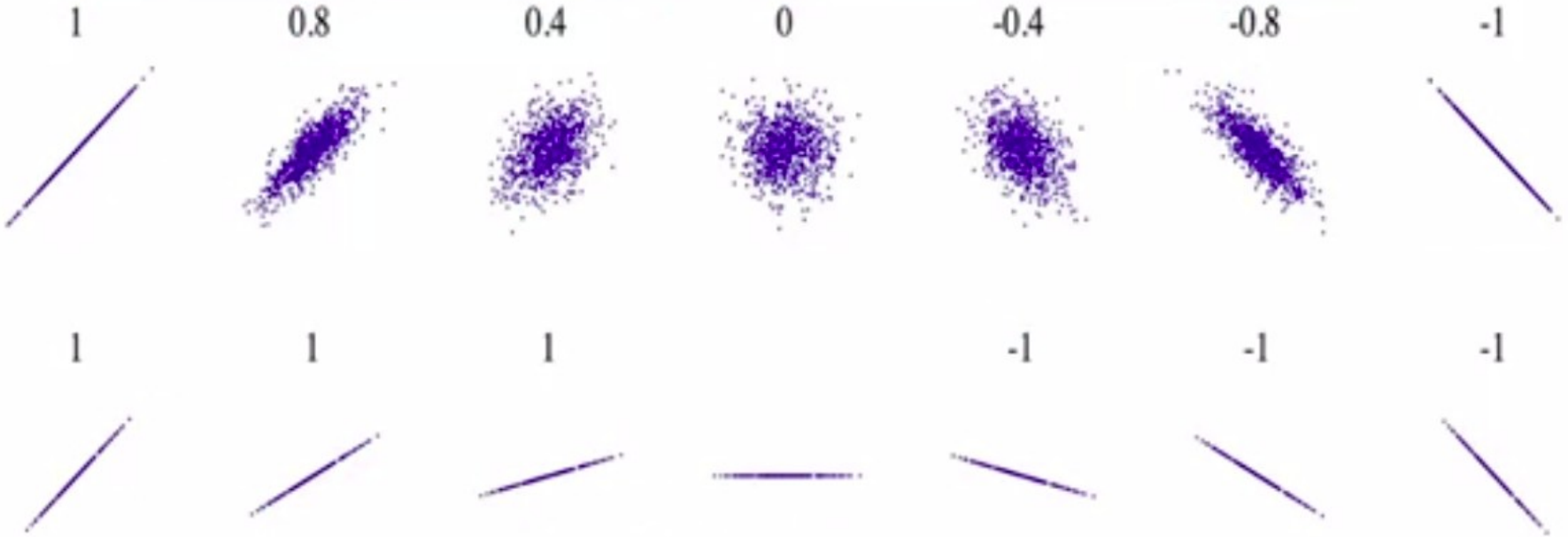
- Measure the strength of the correlation between two features.
  - Correlation coefficient
  - P-value

- Correlation coefficient
  - Close to +1: Large Positive relationship
  - Close to -1: Large Negative relationship
  - Close to 0: No relationship

- P-value
  - P-value < 0.001 **Strong** certainty in the result
  - P-value < 0.05 **Moderate** certainty in the result
  - P-value < 0.1 **Weak** certainty in the result
  - P-value > 0.1 **No** certainty in the result

- Strong Correlation:
  - Correlation coefficient close to 1 or -1
  - P value less than 0.001

Pearson Correlation



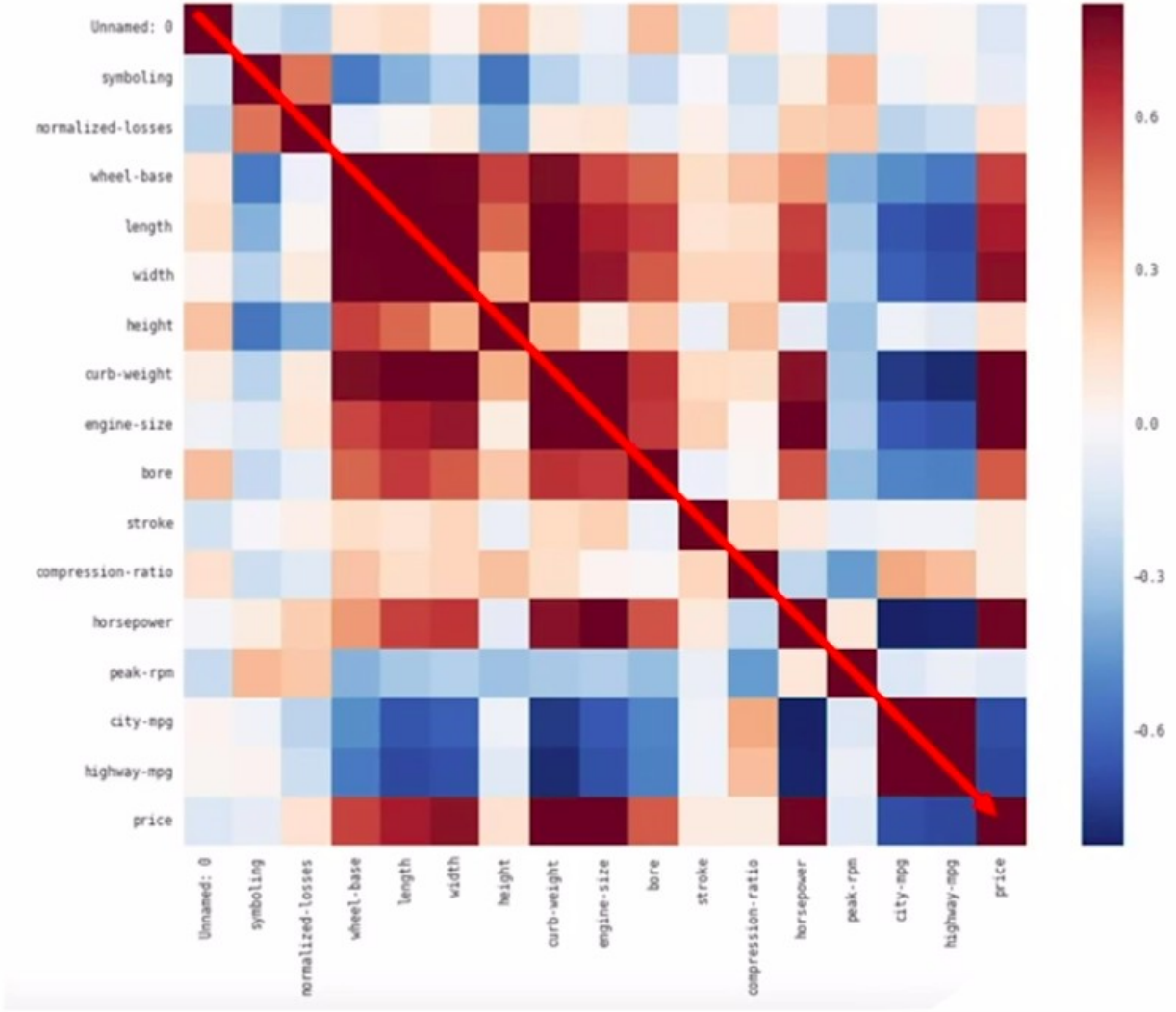
https://en.wikipedia.org/wiki/Correlation_and_dependence

Pearson Correlation- Example

```
pearson_coef, p_value = stats.pearsonr(df['horsepower'], df['price'])
```

- Pearson correlation: 0.81
- P-value : 9.35 e-48

Correlation-Heatmap



Back

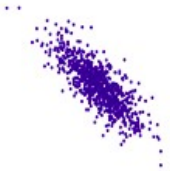
Practice Quiz: Correlation - Statistics

Practice Quiz • 3 min • 1 total point

1. Select the plot with a negative correlation:

1 / 1 point

a



b

Correct

correct

Chi-square: Test for Association

Aije Egwaikhide

Categorical variables

- We use the Chi-square Test for Association (denoted as $\chi 2$)
- The test is intended to test how likely it is that an observed distribution is due to chance.

Chi-Square Test for association

- The Chi-square tests a null hypothesis that the variables are independent.
- The Chi-square does not tell you the type of relationship that exists between both variables; but only that a relationship exists.

Categorical variables

- Is there an association between fuel-type and aspiration?

Observed value

	Standard	Turbo	Total
diesel	7	13	20
gas	161	24	185
Total	168	37	205

Chi-square: Test for Association

Aije Egwaikhide

Categorical variables

- We use the Chi-square Test for Association (denoted as $\chi 2$)
- The test is intended to test how likely it is that an observed distribution is due to chance.

Chi-Square Test for association

- The Chi-square tests a null hypothesis that the variables are independent.
- The Chi-square does not tell you the type of relationship that exists between both variables; but only that a relationship exists.

Categorical variables

- Is there an association between fuel-type and aspiration?

Observed value

	Standard	Turbo	Total
diesel	7	13	20
gas	161	24	185
Total	168	37	205

Categorical variables

- Is there an association between fuel-type and aspiration?

$$\chi 2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Observed value

	Standard	Turbo	Total
diesel	7	13	20
gas	161	24	185
Total	168	37	205

$$\frac{\text{Row total} \quad * \text{Column total}}{\text{Grand total}}$$

Categorical variables

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

	Standard	Turbo	Total
diesel	7	13	20
gas	161	24	185
Total	168	37	205

Observed value

Expected value

aspiration / Fuel-type	Standard	Turbo
Diesel	16.39	
Gas		33.39

$$\frac{\text{Row total} \quad * \text{Column total}}{\text{Grand total}}$$

Categorical variables

$$\chi 2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

	Standard	Turbo	Total
diesel	7	13	20
gas	161	24	185
Total	168	37	205

Observed value

Expected value

aspiration / Fuel-type	Standard	Turbo
Diesel	16.39	3.61
Gas	151.61	33.39

$$\frac{\text{Row total} \quad * \text{Column total}}{\text{Grand total}}$$

Categorical variables

$$\chi 2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Degree of freedom = (row-1)*(column-1)

$$\chi 2 = 29.6$$

Percentage Points of the Chi-Square Distribution

Degrees of Freedom	Probability of a larger value of x^2							
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92

P-value < 0.05, we reject the null hypothesis that the two variables are independent and conclude that there is evidence of association between fuel-type and aspiration.

Categorical variables

$$\chi 2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

	Standard	Turbo	Total
diesel	7	13	20
gas	161	24	185
Total	168	37	205

```
scipy.stats.chi2_contingency(cont_table, correction = True)
```

```
(29.605759385109046,
 5.2947382636786724e-08,
 1,
 array([[ 16.3902439,   3.6097561],
        [151.6097561,  33.3902439]]))
```

Categorical variables

$$\chi 2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

	Standard	Turbo	Total
diesel	7	13	20
gas	161	24	185
Total	168	37	205

```
scipy.stats.chi2_contingency(cont_table, correction = True)
```

```
(29.605759385109046,
 5.2947382636786724e-08,
 1,
 array([[ 16.3902439,   3.6097561],
        [151.6097561,  33.3902439]]))
```

P-value of < 0.05, we reject the null hypothesis that the two variables are independent and conclude that there is evidence of association between fuel-type and aspiration.

Lesson Summary

In this lesson, you have learned how to:

Describe Exploratory Data Analysis: By summarizing the main characteristics of the data and extracting valuable insights.

Compute basic descriptive statistics: Calculate the mean, median, and mode using python and use it as a basis in understanding the distribution of the data.

Create data groups: How and why you put continuous data in groups and how to visualize them.

Define correlation as the linear association between two numerical variables: Use Pearson correlation as a measure of the correlation between two continuous variables

Define the association between two categorical variables: Understand how to find the association of two variables using the Chi-square test for association and how to interpret them.